DOCUMENT RESUME

ED 382 660 TM 023 095

AUTHOR Zwick, Rebecca; And Others
TITLE DIF Analysis for Pretest Items in Computer-Adaptive Testing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-33
PUB DATE May 94
NOTE 53p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Error of Measurement; Estimation (Mathematics); Identification; *Item Bias; Item Response Theory; *Pretests Posttests; Simulation; Test Items
IDENTIFIERS *Mantel Haenszel Procedure; Standardization

ABSTRACT

A simulation study of methods of assessing differential item functioning (DIF) in computer-adaptive tests (CATs) was conducted by Zwick, Thayer, and Wingersky (in press, 1993). Results showed that modified versions of the Mantel-Haenszel and standardization methods work well with CAT data. DIF methods were also investigated for nonadaptive "pretest" items, for which item parameter estimates were assumed unavailable. The pretest DIF statistics were generally well-behaved, but the Mantel-Haenszel DIF statistics tended to have larger standard errors for the pretest items than for the CAT items. The current extension of the earlier work addressed the effect of using alternative matching methods for pretest items. Using a more elegant matching procedure did not lead to a reduction of the Mantel-Haenszel standard errors and produced DIF measures that were nearly identical to those from the earlier study. Further investigation showed that the Mantel-Haenszel standard errors tended to be larger when items were administered to examinees with a wide ability range, whereas the opposite was true of the standard errors of the standardization DIF statistic. Some theoretical findings were obtained that appear to explain this phenomenon. Nine tables and six figures present details of the analyses. (Contains 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 382 660

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

DIF ANALYSIS FOR PRETEST ITEMS IN COMPUTER-ADAPTIVE TESTING

Rebecca Zwick
Dorothy T. Thayer
Marilyn Wingersky

Educational Testing Service
Princeton, New Jersey
May 1994

BEST COPY AVAILABLE

TM023095

02

DIF Analysis for Pretest Items in Computer-Adaptive Testing

Rebecca Zwick, Dorothy Thayer, and Marilyn Wingersky

April 25, 1994

We thank Charlie Lewis, Bob Mislevy, Nancy Petersen, Denny Way, and Neil Dorans for their helpful discussions. We appreciate the sponsorship of the Program Research Planning Committee at ETS.

Copyright © 199-. Educational Testing Service. All rights reserved

Abstract

A simulation study of methods of assessing differential item functioning (DIF) in computer-adaptive tests (CATs) was conducted by Zwick, Thayer and Wingersky (in press; 1993). Results showed that modified versions of the Mantel-Haenszel and standardization methods work well with CAT data. DIF methods were also investigated for nonadaptive "pretest" items, for which item parameter estimates were assumed unavailable. The pretest DIF statistics were generally well-behaved, but the Mantel-Haenszel DIF statistics tended to have larger standard errors for the pretest items than for the CAT items. The current extension of the earlier work addressed the effect of using alternative matching methods for pretest items. Using a more elegant matching procedure did not lead to a reduction of the Mantel-Haenszel standard errors and produced DIF measures that were nearly identical to those from the earlier study. Further investigation showed that the Mantel-Haenszel standard errors tended to be larger when items were administered to examinees with a wide ability range, whereas the opposite was true of the standard errors of the standardization DIF statistic. Some theoretical findings were obtained that appear to explain this phenomenon.

1. Overview:

Zwick, Thayer, and Wingersky (in press; 1993; henceforth referred to as ZTW) conducted an extensive simulation study of DIF methods for CATs. Simulated data were used to investigate the performance of modified versions of the Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) and the standardization method of Dorans and Kulick (1986). Each "examinee" received 25 items out of a 75-item pool. For DIF analysis, examinees were matched on expected true scores based on their CAT responses and on estimated item parameters. Both DIF methods performed well. The CAT-based DIF statistics were highly correlated with DIF statistics based on nonadaptive administration of all 75 pool items and with the true magnitudes of DIF in the simulation. In addition, the across-item means and variances of the DIF statistics were close to their nominal values.

DIF methods were also investigated for 15 nonadaptive "pretest items." Testing programs often include in test forms a set of new items that are being evaluated for further use. In CATs, these pretest items are typically administered nonadaptively to some or all examinees. Because these items are new, item parameter estimates are unavailable. This means that an expected true score cannot be obtained for these items using the conventional formula. The matching variable that was used in ZTW for assessing DIF in these items was the sum of the expected true score on the CAT and the score (0 or 1) on the *studied pretest item* (i.e., the pretest item that was being subjected to DIF analysis).

The pretest DIF statistics were generally well-behaved and had high correlations with the true DIF. A somewhat puzzling result was the finding that the Mantel-Haenszel DIF statistics tended to have larger standard errors for the pretest items than for the CAT items. Some results of Donoghue, Holland, and Thayer (1993) showed that the MH standard error estimates are inflated by the inclusion of the studied item in the matching criterion. Therefore, it seemed possible that the inflated standard errors in ZTW resulted from the method of including the studied item.

If DIF analyses were conducted *after* calibrating the pretest items, it would be possible to include the studied item in the matching variable using an approach that is better grounded in

psychometric theory. Although the necessity of prior calibration would make this *theoretically optimal matching variable* more cumbersome to obtain, the new approach would be a candidate for operational use if it produced results that were significantly better than those in ZTW. Both the optimal matching variable and an approximation to it were investigated in the current study.

Neither the optimal matching variable nor the approximation were found to be substantially different from the matching variable used in ZTW, and DIF statistics obtained using the optimal matching variable had correlations of nearly unity with those obtained in ZTW. The larger standard errors for the *MH D-DIF* statistics were found to be associated with the larger examinee ability range for pretest than for adaptive items.

The steps involved in this study were (a) calibrating the pretest items and reestimating abilities, (b) computing the alternative matching variables, (c) matching examinees on these new scores, (d) recomputing the DIF statistics, and (e) comparing the results to those obtained in ZTW and to the true DIF for the items. These steps are detailed in subsequent sections, following a description of the simulated data.

2. Simulated Examinee Data

DIF analyses are typically based on two groups--the group of primary interest, or *focal* group, and the group to which it is compared, or *reference* group. In this study, the focal group ability distribution was normal, with a mean of -1 and a standard deviation of 1, and the reference group distribution was standard normal. Twenty-five thousand examinees in each of the two groups were included in the study; these were a subset of the 60,000 cases per group that were generated in ZTW. (Section 5.3, below, gives the rationale for using fewer records.) The examinee records from ZTW included the true ability, the expected true score on the CAT, and the responses to the 15 pretest items. The new ability estimates and matching variables needed for the current study, described in sections 3 and 4, were appended to the existing records.

2.1. Generation of Item Responses

The factors that were varied across the items used in ZTW were the item discrimination parameters (*a*) difficulty parameters (*b*), and DIF parameters (*d*). The DIF parameter for item *j* was defined as $d_j = b_{jR} - b_{jF}$, where $b_{jR} = b_j$ is the reference group difficulty and b_{jF} is the focal group difficulty. Therefore, a value of *d* greater than zero implied that an item was easier for the focal group than for the reference group, whereas *d* less than zero implied that the item was harder for the focal group.

Estimates of item parameters and DIF from actual admissions test data were used in determining the values of *a*, *b*, *c*, and *d* for the simulation. (See ZTW for details.) To simplify the simulation, the guessing parameter, c_j, was set equal to .15 for all items. Item responses were generated with the three-parameter logistic (3PL) item response function (Birnbaum, 1968), using the true item and ability parameters.

2.1.1. CAT Item Pools

In ZTW, responses to three different CAT pools of 75 items were simulated. In Pool 1, the items had no DIF; in Pool 2, the items had DIF that was uncorrelated with item difficulty; and in Pool 3, the items had DIF that was correlated with item difficulty. The same set of 15 pretest items accompanied each pool. Although each examinee's responses to CAT items were used in computing the matching variable for that examinee, DIF results for the pretest items were found to be nearly identical across pools. Therefore, the present study used only one CAT pool: Pool 2. The included values of *a*, *b*, and *d* for Pool 2 items were as follows:

a: .74, 1,

b: -1.95, -1.3, -.65, 0 .65, 1.3, 1.95, and

d: -.70, -.35, 0, .35, .70.

The CAT simulation was designed as a simplified version of actual CATs being developed at ETS. The CAT algorithm, originally developed for the ZTW study, selects as the next item to be administered the most informative item at the maximum likelihood estimate (MLE) of ability

computed from the items already administered. The item parameter estimates used for computing item information and ability estimates were obtained through an analog to a paper-and-pencil test administration. A sample of 2,000 reference group examinees were "administered" all 75 items, and the LOGIST program (Wingersky, 1983; Wingersky, Patrick, & Lord, 1988) was used to estimate the *a*, *b*, and *c* parameters. Because 2,000 is a typical sample size for such calibrations, this approach allowed for the incorporation of a realistic amount of estimation error. The true and estimated *a*, *b*, and *c* parameters, along with the true *d* parameters, are given in ZTW (see listing for Pool 2).

2.1.2. Pretest Items

All 15 pretest items had a true *a* value of 1 and a true *c* value of .15. Three values of *b* (1.3, 0, and 1.3) were crossed with five values of *d* (-.7, -.35, 0, .35, and .7). The true and estimated *a*, *b*, and *c* parameters, along with the true *d* parameters, are given in Table 1.

Insert Table 1 about here.

---

3. Pretest Item Calibration and Ability Estimation

The item parameter estimates for the pretest items, which were needed to compute the new matching variables, were obtained by including the 15 pretest items in a single calibration run with the 75 CAT items. Parameter estimates for the 75 CAT pool items were fixed at the values previously obtained in ZTW. The calibration sample originally used in estimating the CAT item parameters (see section 2.1.1) was used. The procedure was intended to parallel a paper-and-pencil calibration in that all examinees had responses to all 90 items.[1]

For each pretest item, each examinee's ability was reestimated using the 25 CAT item responses, as well as the response to the studied pretest item. This resulted in 15 new ability estimates for each examinee.

4. Matching Variables for Pretest DIF Analysis

In ZTW, the matching variable for the DIF analysis of the CAT-administered items was obtained by (1) getting the examinee's MLE of ability, based on the responses to the 25 CAT items and (2) using this MLE, along with the estimated item parameters, to compute an expected true score for the entire item pool by summing the 75 values of the estimated item response functions. The matching variable for the DIF analysis of pretest items was obtained by adding the score on the studied pretest item to the CAT matching variable. That is, the old matching variable for the *i*th pretest item was

$$\sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_{CAT}) + X_i, \qquad (1)$$

where $\hat{P}_j(\cdot)$ is an estimate of the 3PL function for the *j*th item, $\hat{\theta}_{CAT}$ is the MLE of ability based on the CAT items, and X_i is the score on the *i*th pretest item.

The *theoretically optimal matching variable* was obtained as follows:

$$\sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_i) + \hat{P}_i(\hat{\theta}_i) \qquad (2)$$

where $\hat{\theta}_i$ is the MLE of ability based on the CAT and the *i*th pretest item. This matching variable is the expected true score on the CAT pool and the *i*th pretest item, with the ability estimate based on 26 item responses.

The *approximation to the optimal matching variable* was

$$\sum_{j=1}^{75} \hat{P}_j(\hat{\theta}_{CAT}) + \hat{P}_i(\hat{\theta}_{CAT}) \qquad (3)$$

In this approximation, the MLE of ability is based on the CAT only, as in equation 1, but the expected true score is defined in terms of the 75 CAT pool items and the studied pretest item, as in equation 2. An interesting feature of this approximation is that, although it uses the item parameter estimate for the studied pretest item, it does not use the examinee's response to that item.

The matching variables in equations 1-3 depend on the parameter estimates or responses for the *studied* pretest item, but do not depend on the estimates or responses for the remaining 14 pretest items. Therefore, in the pretest DIF analyses in ZTW and in the present study, the matching variable differed across pretest items.

4.1. Distributions of Residuals for the Three Matching Variables

Table 2 gives descriptive information on the distributions of the residuals for the matching variables in equations 1-3. Results are given separately for the reference and focal groups, based on samples of 1,000. The residual for an examinee was obtained by subtracting the examinee's true score (based on the generating abilities and item parameters) from the value of the matching variable. The true score for the *i*th pretest item was defined as

$$\sum_{j=1}^{75} P_j(\theta) + P_i(\theta). \qquad (4)$$

Insert Table 2 about here.

The standard deviations of the true scores were approximately 13 for the reference group and 12 for the focal group. The median residuals in Table 2 are very small with respect to this standard deviation and show little variation across matching variables. The largest departure from zero occurred for the optimal matching variable in the focal group results for Item 1, with a median residual of -.45. The approximation to the optimal matching variable showed the least variation across items.

As in ZTW, nearly all median residuals were negative. An investigation of this phenomenon in ZTW led to the conclusion that the obtained set of item parameter estimates led to a downward bias in the ability estimates. The matching variables used in the present study depend in large part on the responses and parameter estimates for the CAT items; therefore, it is not surprising that negative residuals were preponderant in this study as well.

5. DIF methods

This section describes the two DIF methods and the procedure used to estimate the behavior of DIF statistics under various sample size conditions. In both the MH and standardization methods, examinees are first grouped on the basis of a matching variable that is intended to be a measure of ability in the area of interest. In many DIF applications, the matching variable is the total score on the test in which the studied item is embedded. In the current study, the matching variable was given by equation 2. Examinees whose values on the matching variable fell in the same one-unit intervals were considered to be matched. Because the DIF results obtained with the optimal matching variable were not found to be superior to those obtained using the simple matching approach in ZTW, no DIF analyses were conducted using the approximation to the optimal matching variable (equation 3).

For the DIF analyses in the current study, each of the 15 pretest items in turn played the role of the studied item. The score on the studied item, group membership, and the value of the matching variable for each examinee define a 2 x 2 x K cross-classification of examinee data, where K is the number of levels of the matching variable. Assume that there are T_k examinees at the kth level of the matching variable. Of these, n_{Rk} are in the reference group and n_{Fk} are in the focal group. Of the n_{Rk} reference group members, A_k answered the studied item correctly while B_k did not. Similarly C_k of the n_{Fk} matched focal group members answered the studied item correctly, whereas D_k did not.

5.1. Mantel-Haenszel DIF Analysis

The MH measure of DIF is

$$MH\ D\text{-}DIF = -2.35 \ln(\hat{\alpha}_{MH}) \qquad (5)$$

where $\hat{\alpha}_{MH}$ is the Mantel-Haenszel conditional odds-ratio estimator given by

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k} \qquad (6)$$

In equation 5, the transformation of $\hat{\alpha}_{MH}$ places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The effect of the minus sign is to make *MH D-DIF* negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. An estimated standard error for *MH D-DIF*, based on work by Robins, Breslow and Greenland (1986) and Phillips and Holland (1987), is given in Holland and Thayer (1988). It is

$$SE(MH\ D-DIF) = 2.35\sqrt{Var(\ln(\hat{\alpha}_{MH}))} \qquad (7)$$

where $Var(\ln(\hat{\alpha}_{MH}))$ is estimated by

$$\frac{\sum_k U_k V_k / T_k^2}{2\,(\sum_k A_k D_k / T_k)^2}, \qquad (8)$$

where $U_k = (A_k\ D_k) + \hat{\alpha}_{MH}(B_k\ C_k)$ and $V_k = (A_k + D_k) + \hat{\alpha}_{MH}(B_k + C_k)$.

The Mantel-Haenszel chi-square test of the null hypothesis of no DIF was not examined in this research.

5.2. Standardization Analysis

The standardization DIF measure, developed by Dorans and Kulick (1986), is

$$STD\ P-DIF = \hat{p}_F - \tilde{p}_R \tag{9}$$

where $\hat{p}_F$ is the proportion in the focal group who get the studied item correct, and $\tilde{p}_R$ is an adjusted proportion correct on the item for the reference group, defined as

$$\tilde{p}_R = \sum_k \left(\frac{A_k}{n_{Rk}}\right) \frac{n_{Fk}}{n_F} \tag{10}$$

where $n_F = \sum_k n_{Fk}$ is the total number of examinees in the focal group.[2] One interpretation of $\tilde{p}_R$ is that it is the proportion of reference group examinees who would have got the studied item right if the distribution of the matching variable in the reference group had been the same as in the focal group.

The estimated standard error for $STD\ P-DIF$ is given by the formula

$$SE(STD\ P-DIF) = \sqrt{\sigma_F^2 + \sigma_R^2} \tag{11}$$

where

$$\sigma_F^2 = \frac{1}{n_F}\,\hat{p}_F\,(1-\hat{p}_F) \tag{12}$$

and

$$\sigma_R^2 = \frac{1}{n_F^2} \sum_k \frac{n_{Fk}^2 A_k B_k}{n_{Rk}^3}. \qquad (13)$$

5.3. Definition of Sample Size Conditions

Three sample size conditions were of interest: $n_R = 500$, $n_F = 500$; $n_R = 900$, $n_F = 100$; and $n_R = 500$, $n_F = 100$, where n_R and n_F are the sample sizes for the reference and focal groups, respectively. These sample size conditions were chosen to be similar to those that are expected to occur in ETS analyses of pretest items that accompany CATs. The first two sample size conditions were also used in ZTW.

The difficulty of defining sample sizes in a CAT simulation was addressed in ZTW. If groups of a fixed sample size had been drawn and the CAT administered, the sample sizes *per item* would have had a huge range. Because the goal was to investigate the behavior of DIF statistics for specific sample sizes, it would not have been useful simply to analyze the available data for each item. After considering several other approaches, including resampling techniques and multiple replications, the *expected table* (ET) method was adopted:[3] First, within each simulation condition, item response data were generated for 60,000 examinees per group. Each examinee received 25 of the 75 CAT pool items. For each of the 75 items, all the available CAT data (out of a maximum of 60,000 responses per group) were then used to form the 2 (item responses) x 2 (groups) x K (levels of the matching variable) contingency table needed for DIF analysis. (The response frequency per item ranged from about 1,800 to about 34,000.) The table frequencies were then converted to proportions of the total number of observations for the group in question. Using these proportions as estimates of the population probabilities associated with the 4 x K cells for the relevant configuration of conditions, expected tables for the target sample sizes were obtained by multiplying the probability estimates for focal group cells by the desired focal group sample size and then doing the same for the reference group cells. Next, DIF statistics and

standard errors were computed, based on the expected tables. A simple example of the ET approach, which originally appeared in ZTW, is given in the Appendix.

Though it produces only a single estimate, the ET approach can provide a relatively precise idea of the behavior of the *MH D-DIF* statistic. A supplementary study comparing the ET method to an estimation procedure based on multiple replications (as in a typical simulation study) appears in ZTW. The comparison was based on items for which 60,000 responses per population group were available. The ET method was found to give results similar to those of the replication-based approach. For the items that were studied, the ET-estimated *MH D-DIF* was determined to be as precise as an average over 316 replications of the *MH D-DIF* statistic based on the target sample sizes. Another advantage of the ET approach is that, once the 2 x 2 x *K* probability tables have been created, DIF results can be generated easily for any target sample size, facilitating further research.

The advantages of the ET method are less clear for the pretest items, for which data are available for all simulated examinees, than for CAT items. To facilitate comparisons with the CAT items, however, the ET method was used for the pretest items in ZTW. The current study used the ET method so that pretest item results could be compared to those in ZTW. Examination of ZTW results showed that increasing the number of examinees per group beyond 25,000 led to very little gain in the precision of the ET estimates; therefore, only 25,000 of the 60,000 available records for each group were included in the current study.

Except where noted, the values of *MH D-DIF*, *SE(MH D-DIF)*, *STD P-DIF*, and *SE(STD P-DIF)* in this report were computed from expected tables using equations 5 through 13. Note that *SE(MH D-DIF)* and *SE(STD P-DIF)* are not indexes of the error associated with the *estimation* of *MH D-DIF* and *STD P-DIF*. Instead, these standard errors closely approximate the values of *SE(MH D-DIF)* and *SE(STD P-DIF)* that would be obtained using actual samples of the target sizes. The appropriate formulas for the standard errors of the ET estimates of *MH D-DIF* and *STD P-DIF*, which reflect the degree of precision with which the population DIF values

are estimated using the ET approach, are given in ZTW. In the present study, these standard errors of the estimate ranged from .05 to .08 for *MH D-DIF* and from .004 to .005 for *STD P-DIF*.

6. DIF Results

DIF analyses were conducted on the 15 pretest items for each of the three sample size conditions, using the optimal matching variable. Results were compared to those in ZTW and to the true DIF for the item. For purposes of these analyses, the true DIF of an item was defined as the generating value of *ad* for that item (see Table 2), based on the following rationale: Under certain Rasch model conditions, the *MH D-DIF* statistic provides an estimate of 4*ad*. (See Donoghue, Holland and Thayer, 1993, who based their result on the work of Holland and Thayer, 1988.) The assumptions under which this finding holds are that (1) within each of the groups (reference and focal), the item response functions follow the Rasch model (obtained by setting c_j = 0 for all items j and $a_j \equiv a$ for all items j) , (2) the matching variable is the number-right score based on all items, *including* the item under analysis, referred to as the *studied item*, and (3) the items have the same item response functions for the reference and focal groups (i.e., $b_{jR} = b_{jF} \equiv b_j$), with the possible exception of the studied item. When conditions (1)-(3) do not hold, the population odds ratios will not, in general, be constant across number-right score levels (see Zwick, 1990). Therefore, for the 3PL model, it is not possible to derive a general expression for the quantity estimated by the *MH D-DIF* statistic. However, empirical analyses showed that, in the ZTW study, *MH D-DIF* was approximately equal to 3*ad*.

Table 3 gives the means and standard deviations across the 15 items of *MH D-DIF*, *SE(MH D-DIF)*, *STD P-DIF*, and *SE(STD P-DIF)*. These summaries are given for the ZTW results (two sample size conditions) and the new results (three sample size conditions). Two artifacts of the simulation procedures need to be considered when interpreting these results. First, *within each study*, results for the various sample size conditions are based on the same ET probability tables. Because of this, the *MH D-DIF* measures within each of the two studies are highly correlated across sample size conditions. The ET-estimated *STD P-DIF* statistic is

invariant over target sample sizes. The primary purpose of including more than one sample size condition was to study the behavior of the standard errors of the DIF measures under various sample sizes. Second, as described in section 5.3, the ET probability tables for the new study are based on a subset of the simulated examinees from the ZTW study. The ET tables are not the same as in ZTW because only a subset of the examinees was used and because the examinees were stratified using a different matching variable.

Insert Table 3 about here.

The mean across items of the DIF values used in data generation was zero; ideally, the mean of the DIF statistics would also be zero. The standard deviation of the generating DIF values was about .5; the *MH D-DIF* statistics were expected to have a standard deviation roughly three times that large. Results from the old and new analyses look very similar, with the mean DIF statistics for the new analyses departing slightly more from zero than the means for the old analyses. (When DIF analyses are conducted in which the matching variable is the number-right score, including the studied item, the *STD P-DIF* statistics are constrained to sum to zero across items; the *MH D-DIF* statistics are also constrained to sum to approximately zero. This constraint does not apply in the present analyses; that is, the closeness to zero of the means of Table 3 was not a foregone conclusion.) In both the new and old analyses, the across-item variation of the *MH D-DIF* statistics was somewhat smaller than expected.

The stability of the DIF statistics depends heavily on the minimum of the two sample sizes; therefore the variation of the standard errors across sample size conditions was as expected: The standard errors were considerably larger for $n_R = 900$, $n_F = 100$ than for $n_R = 500$, $n_F = 500$, and only slightly larger for $n_R = 500$, $n_F = 100$ than for $n_R = 900$, $n_F = 100$. See section 6.2 for further discussion of these results.

There was also a tendency for standard errors to be slightly larger for the new analyses. Figures 1 and 2 show the standard errors for the new and old analyses for *MH D-DIF* and *STD*

P-DIF, respectively for n_R= 500, n_F = 500. On nearly every item, the standard errors of both of these statistics were slightly larger for the new than for the old analyses. This suggests that better matching may have been achieved using the old matching variable (equation 1).[4]

Insert Figures 1-2 about here.

Tables 4 and 5 give the intercorrelations of the old and new DIF measures and true DIF for the n_R = 500, n_F = 500 and the n_R = 900, n_F = 100 sample size conditions, respectively. (Reliabilities were computed for each DIF statistic using the methodology described in ZTW. Because they were all at least .997, reliability-corrected results were essentially identical to uncorrected results.) For both sample size conditions, the intercorrelations among the DIF statistics were nearly unity. The correlations with true DIF (*ad*) ranged from .91 to .95; these correlations were .01 higher for the new than for the old analyses.

Insert Tables 4-5 about here.

6.1 Expected Percent of "C" Results

ETS has a system for categorizing the severity of DIF based on MH results. According to this classification scheme, a "C" categorization, which represents extreme DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at α = .05). A "B" categorization, which indicates moderate DIF, requires that *MH D-DIF* be significantly different from zero (at α = .05) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to be free of DIF. Items that fall in the C category are typically eliminated from tests or subjected to further scrutiny.

Because the ET-estimated DIF statistics were based on a total of 50,000 observations in the present study, it is reasonable to assume that they provide precise estimates of the population mean and standard deviation of the theoretical distribution of *MH D-DIF* for the relevant configuration of item properties and simulation conditions. This is supported by analyses reported in ZTW. If it is assumed that *MH D-DIF* statistics for this configuration follow a normal distribution with this mean and standard deviation, percentiles of the theoretical distribution of *MH D-DIF* can be obtained. These percentiles can then be used to estimate the percent of times such an item will be classified as an A, B, or C. This is an alternative way of providing information about the sampling variation of the *MH D-DIF* statistic.

Based on the ETS DIF rules, an algorithm was developed for estimating these percents, to be applied separately to each item in each condition. The algorithm was tested and found to work well with data from two simulation studies. Details are given in ZTW.

In the 3PL model, the determination of which items are *nominally* A, B, and C items is not straightforward. Based on the empirical finding that *MH D-DIF* was approximately equal to 3*ad* in the conditions investigated in ZTW, Items 1, 5, 6, 10, 11, and 15, which have $ad = \pm.70$, can be considered to be nominal C's; Items 2, 4, 7, 9, 12, and 14, which have $ad = \pm.35$, would be nominal B's; and Items 3, 8, and 13, which have $ad = 0$ would be nominal A's. The same categorization would result from application of the theoretical finding that *MH D-DIF* is an estimate of 4*ad* under certain Rasch model conditions. (See ZTW for further discussion of these issues.)

Table 6 gives the expected percent of C results for the 15 pretest items for the new and old matching variables for each sample size condition. The nominal status of the item is also given, with a "minus" sign indicating items that are easier for the reference group and a "plus" sign indicating items that are easier for the focal group. As expected, the likelihood of detecting C items is greater when *SE(MH D-DIF)* is smaller (see section 6). Results tended to be similar for the new and old matching variables. The exceptions were two nominal C items: Item 10 for $n_R = 500$, $n_F = 500$, in which the new analysis was better able to detect DIF and Item 1 for $n_R = 900$,

n_F = 100, in which the old analysis was more sensitive. For both the new and old analyses, DIF became harder to detect as items became more difficult. This is most obvious in the case of Item 11, a nominal C item with b = 1.30, for which the expected percent of C results never exceeded 1.4. This phenomenon, noted by Donoghue, Holland, and Thayer (1993), occurs in simulations in which the guessing parameter c is constrained to be the same in the reference and focal groups. The more difficult the item, the closer the probability of correct response is to the guessing value, and the harder the groups are to differentiate.

Insert Table 6 about here.

6.2. DIF Results for Smaller Sample Sizes

A secondary question in this study concerned the functioning of the *MH D-DIF* statistic with n_R = 500, n_F = 100. It is anticipated that sample sizes this small will be encountered in pretest DIF analyses at ETS. Specifically, there was an interest in whether the standard errors of the DIF statistics would be much larger than for the n_R = 900, n_F = 100 condition included in ZTW.

Table 7 gives a comparison of the distribution across items of the standard errors of the DIF statistics for the new analyses for n_R = 900, n_F = 100 and n_R = 500, n_F = 100. For both the *MH D-DIF* and *STD P-DIF* statistics, the median standard error for n_R = 500, n_F = 100 was larger than the median for n_R = 900, n_F = 100 by a factor of 1.06.

Note that by using the estimated value of the standard error for a particular pair of sample sizes (say, n_{R1}, n_{F1}) as a baseline, the standard error of *MH D-DIF* for another sample size pair (say, n_{R2}, n_{F2}) can be predicted accurately using the ratio of the harmonic means of the sample size pairs. More specifically, $SE(n_{R2}, n_{F2})$ can be well predicted by multiplying $SE(n_{R1}, n_{F1})$ by $\sqrt{\frac{h(n_{R1}, n_{F1})}{h(n_{R2}, n_{F2})}}$, where h(.) denotes the harmonic mean. In the present case, the multiplication factor obtained through the simulation (1.06) is approximately equal to that which would be

obtained analytically by examining the ratio of the harmonic mean of the sample sizes for the two conditions (1.04).

A point to keep in mind when evaluating the results of Table 7 is that DIF standard errors for both $n_R = 900$, $n_F = 100$ and $n_R = 500$, $n_F = 100$ are quite large relative to the magnitudes of DIF that are sometimes thought to be of interest. As shown in Table 6, the detection of C items is substantially impaired when the smaller of the two sample sizes is 100 rather than 500.

Insert Table 7 about here.

---

7. Further investigation of Standard Errors for Pretest DIF Statistics

The ZTW study showed that *SE(MH D-DIF)* tended to be larger for pretest items than for CAT items, with a range of 0.4 - 0.6 for the $n_R = 500$, $n_F = 500$ condition and 0.6 - 1.1 for the $n_R = 900$, $n_F = 100$ condition, compared to 0.3 - 0.4 and 0.5 - 0.7, respectively, for the two sample size conditions in the CAT. The values of *SE(STD P-DIF)*, however, were slightly *smaller* for pretest items than for CAT items.

The findings of the present study show clearly that the large values of *SE(MH D-DIF)* did not result from the ad hoc procedure used in ZTW for defining the matching variable. As shown in Figure 1, standard errors were slightly larger when the theoretically optimal matching variable was used. Other factors were therefore investigated to determine the cause of the seemingly inflated MH standard errors. Two questions that were addressed were:

How do the standard errors compare for CAT and pretest items if the matching variable is held constant?

Do the standard errors of *MH D-DIF* and *STD P-DIF* accurately reflect the empirical variation of these DIF indexes?

7. 1. How do the standard errors compare for CAT and pretest items if the matching variable is held constant?

Figures 3 and 4 show the results of computing two kinds of *MH D-DIF* statistics for 71 items administered in the ZTW CAT study (Pool 3). (Four of 75 CAT pool items were never administered; see ZTW.) The values plotted along the horizontal axis are based on a nonadaptive administration to 900 reference and 100 focal group examinees. The matching variable was the expected true score on the 75 CAT pool items, with the ability estimate based on responses to all 75 items. The values plotted along the vertical axis are based on only the examinees who received the item in a CAT administration. The matching variable, however, was the same as that used for the nonadaptive DIF statistics. In practice, it would, of course, be impossible to compute this matching variable for examinees who received only 25 CAT items. The reason for doing so in this analysis was to eliminate any possible effect of the matching variable on the DIF statistics. The CAT DIF results for the $n_R = 900$, $n_F = 100$ sample size condition were obtained using the ET method.

Insert Figures 3-4 about here.

Figure 3 shows that the *MH D-DIF* statistics are clustered around the 45-degree line; there were no systematic differences for the two types of administration. Figure 4, however, shows that the standard errors were dramatically different; for almost every item, they were smaller for the CAT data.

To determine whether these findings were related to the particular CAT algorithm used in this study, another analysis was conducted, comparing what might be termed a pseudo-CAT administration to nonadaptive administration. For the pseudo-CAT, examinees were eliminated from the DIF analysis of a particular item if their abilities departed substantially from the estimated item difficulty. Specifically, if the examinee's ability was less than $\hat{b}$ - 1.25 or greater than $\hat{b}$ + 1.25, the examinee was eliminated. The ET method was applied to both the selected and

unselected data, with target sample sizes of $nR = 900$, $nF = 100$. Results were very similar to those obtained in Figures 3 and 4, suggesting that this phenomenon was associated with the ability range of the examinees, but was not unique to the ZTW CAT algorithm. In a later study of DIF analysis for pretest items, Way (1994) obtained a similar result.

To determine whether this result was related to the use of item response theory for ability estimation, the CAT DIF statistics were also plotted against those obtained through nonadaptive administration, with number-right score as a matching variable. Again, results were very similar to those in Figures 3 and 4.

Figures 5 and 6 show corresponding results for the *STD P-DIF* statistic and its standard error. The *STD P-DIF* values are clustered around the 45-degree line, but, as in ZTW, the standard errors showed a tendency to be *smaller* for the nonadaptive administration.

Insert Figures 5-6 about here.

7.2. Do the standard errors of MH D-DIF and STD P-DIF accurately reflect the empirical variation of these DIF indexes?

The ZTW study included a comparison of the ET method to an estimation procedure that involved averaging over 66 replications. The DIF analyses for this investigation were conducted on the 15 pretest items. For each item, the comparison produced three estimates of the variation of the *MH D-DIF* and *STD P-DIF* statistics: the empirical variation of *MH D-DIF* or *STD P-DIF* across the 66 replications, the average over 66 replications of the ordinary standard errors, computed using equations 7-8 and 11-13, and the ET standard error, computed on the expected tables using these same equations. For the ET estimates, the total sample size was 60,000 per group, and the target sample sizes were $nR = 900$ and $nF = 100$. For the replication-based approach, actual samples of 900 reference and 100 focal examinees were used.

Table 8 shows the Pearson correlations (across the 15 items) among these three estimates of the variability of *MH D-DIF* and the corresponding three estimates for *STD P-DIF*. Several aspects of these results are worthy of note: The ET standard errors were highly correlated with the mean SE across replications (.99 for *MH D-DIF*; .98 for *STD P-DIF*). Also, the ET standard errors had moderately high correlations with the empirical standard deviation of the DIF indexes (.81 for *MH D-DIF* and for *STD P-DIF*). All cross-correlations between the MH-based variables and the standardization-based variables were negative and some were substantial in magnitude. The empirical standard deviations of *MH D-DIF* and *STD P-DIF* had a negative correlation, albeit a small one (-.14).

Insert Table 8 about here.

In discussing the somewhat inferior performance of *SE(STD P-DIF)* relative to *SE(MH D-DIF)* in another simulation study, Donoghue, Holland, and Thayer (1993) made note of "the large amount of work that went into the development of a useful standard error for the Mantel-Haenszel log-odds-ratio estimator...In contrast, [the formula for *SE(STD P-DIF)* is] based on simple asymptotic approximations that are suspect when the number of cases in each 2 x 2 table is small" (p. 161, 163). The correlational results in Table 8, however, show that the standard errors of both *MH D-DIF* and *STD P-DIF* paralleled the corresponding empirical variability quite well.

Table 9 gives the medians and ranges across the 15 items for the three types of standard error estimates. For *MH D-DIF*, the ET standard error tended to be slightly smaller than the empirical standard deviation, while for *STD P-DIF*, the ET standard error tended to be slightly larger. For both *MH D-DIF* and *STD P-DIF*, the range over items of the ET standard error was slightly smaller than the range of the empirical standard deviation. For both the MH and standardization statistics, the mean standard error across replications tended to be slightly larger than the empirical standard deviation.

Insert Table 9 about here.

7.3 Analytical Perspective on the Standard Errors of DIF Statistics

The standard error findings described in this paper are consistent with the analytical results in Zwick (1993). Those results are briefly summarized here.

The *MH D-DIF* and *STD P-DIF* statistics use different metrics to compare the item performance of two population groups. In *MH D-DIF*, the *ratio of two odds* is examined, whereas the *STD P-DIF* statistic is based on the *difference between two proportions*. Both statistics are, of course, more complicated in that they involve conditioning on a matching variable and include special weighting functions. To understand the variability of DIF statistics, it is useful to start by considering a simpler but analogous problem: Suppose that there is no stratification (i.e., $K = 1$) and we are concerned with a sample proportion, $\hat{p}$, for a single population. We can compare the variance of this proportion to the variance of the logit of the proportion. The variance of a sample proportion, under the binomial model, is $n^{-1}\pi(1-\pi)$, where π is the population proportion. The asymptotic variance of logit $(\hat{p}) = \ln[\hat{p}/(1-\hat{p})]$ is $[n\pi(1-\pi)]^{-1}$ (Agresti, 1990). These variances are estimated by substituting $\hat{p}$ for π.

Two related facts are worthy of note here. First, the variance of $\hat{p}$ and the variance of logit($\hat{p}$) are inversely related. Second, when $\hat{p} = .5$, the estimated variance of $\hat{p}$ is maximized while the estimated variance of logit($\hat{p}$) is minimized. Zwick (1993) shows that, when the Rasch model holds and there is no DIF, the exact variance of a Taylor series approximation to the MH log odds ratio is equal to

$$\mathrm{Var}\left(\ln\left(\hat{\alpha}_{\mathrm{MH}}\right)\right) = \left\{\sum_{k=1}^{K}\left(\frac{n_{Rk}n_{Fk}}{n_{Rk}+n_{Fk}}\right)\pi_k\left(1-\pi_k\right)\right\}^{-1},$$

where π_k is the probability of a correct answer in stratum k. This expression is obviously similar to the expression for the asymptotic variance of a logit. Each of the K terms will be minimized if the stratum probability of a correct response on the item is equal to .5; the variance becomes large under departures from this condition. The $\hat{P} = .5$ condition is similar to the situation that occurs in a CAT, when items are administered to examinees of an appropriate, and thus rather narrow, ability range. Large departures from this condition occur on the pretest. These findings appear to explain both the differences between the CAT and the pretest standard errors and the differences between the *MH D-DIF* and *STD P-DIF* standard errors.

8. Summary:

This study yielded several practical findings about the analysis of DIF in CAT and CAT pretest items:

- An elaborated method of obtaining the matching variable for pretest items, which involved prior calibration of these items, did not produce results superior to those obtained with the simple matching variable presented in Zwick, Thayer and Wingersky (1993). The new DIF statistics had correlations of nearly unity with the previous results, and the standard errors were slightly larger in the new analysis. Therefore, the simple approach presented earlier is recommended.

- The standard errors of *MH D-DIF* and *STD P-DIF* for $n_R = 500$, $n_F = 100$ did not greatly exceed those for $n_R = 900$, $n_F = 100$. The increases in the standard errors relative to the $n_R = 900$, $n_F = 100$ conditions were consistent with theoretical estimates based on the reference and focal group sample sizes. Using the simulation results obtained in this study and in ZTW as a starting point, it should be possible to use analytical means to predict the standard errors for various sample size combinations.

• Both empirical and theoretical results indicated that the standard errors for *MH D-DIF* and *STD P-DIF* are related to the proportion correct on the item. When the proportions correct within the score strata are close to .5, which is consistent with CAT administration, *SE(MH D-DIF)* tends to be small, while *SE(STD P-DIF)* tends to be large. Departures from this condition, which are sometimes extreme for the pretest items, tend to increase *SE(MH D-DIF)* and decrease *SE(STD P-DIF)*.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wily.

Birnbaum, A. (1968). Some latent trait models. In Lord, F. & Novick, M. (1968), *Statistical theories of mental test scores*, pp. 397-424. Reading, MA: Addison-Wesley.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (eds.) *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. ETS Research Report No. RR 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

McHale, F., Dorans, N., Holland, P., & Petersen, N. (May 2, 1988). Specifications for standardized percent correct and distractor analysis (IANA80 and IANA82). Technical memorandum, Educational Testing Service.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425-431.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*, 311-323.

Way, W. D. (1994). *A simulation study of the Mantel-Haenszel procedure for detecting DIF with the NCLEX using CAT.* Technical report, Educational Testing Service.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (ed.), *Applications of item response theory.* Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST user's guide: LOGIST Version 6.00.* Princeton, NJ: Educational Testing Service.

Zwick (1993). *The effect of probability of correct response on the variability of measures of differential item functioning.* Submitted to ETS Research Report series.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185-197.

Zwick, R., Thayer, D. T., & Wingersky, M. (in press). A simulation study of methods for assessing differential item functioning in computer-adaptive tests. *Applied Psychological Measurement.*

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* ETS Research Report 93-11. Princeton, NJ: ETS.

Footnotes

1 A second calibration method was initially considered which featured item administration and calibration procedures more similar to those that have been developed for ETS CATs. In addition to the steps involved in the implemented pretest calibration, this method required that a set of "anchor items" be used to link the pretest items to an existing scale. Because of errors of equating, this method would have necessarily led to results inferior to those obtained with the optimal matching variable. Because the simple method used in ZTW was found to produce essentially the same results as those obtained with the optimal matching variable, there was no reason to pursue additional methods that required calibration of the pretest items.

2 When n_{Rk} is equal to zero, both $\bar{p}_R$ and σ_R^2 are undefined. When this occurs, the standard ETS DIF software implements an imputation procedure proposed by Holland, (McHale, Dorans, Holland & Petersen, 1988). A modification of this procedure which takes into account the special nature of the CAT-based analyses was used in this study.

3 This approach was proposed by Charles Lewis.

4. The values of *SE(MH D-DIF)* and *SE(STD P-DIF)* depend on the target sample sizes (e.g., $n_R = 900$, $n_F = 100$); they do not depend on the sample sizes on which the ET estimates are based (in this case, 25,000 per group). The ET sample sizes determine the precision with which the target tables reflect the generating probabilities, but the degree of this precision does not affect the magnitude of the statistics that are computed using the target tables. An empirical verification of this was obtained in a related study, in which DIF results were obtained for ET sample sizes of 25,000 per group and for ET sample sizes of 60,000 per group. Although the standard errors of the ET estimates (not presented in the current study) varied across the two analyses, the values of *SE(MH D-DIF)* and *SE(STD P-DIF)* for the target tables were nearly always identical to two decimal places across analyses.

Appendix: Example of the Expected Table (ET) Approach to Estimation of DIF Statistics

Consider the following hypothetical results for a single item, assuming that there are only two levels of the matching variable. The first step is to use all the data available for the item to construct a 2 x 2 x 2 frequency table (because $K = 2$ here). Then the cell frequencies for the reference group are divided by the total number of reference group examinees and the frequencies for the focal group are divided by the total number of focal group examinees, producing the following 2 x 2 x 2 table of probabilities:

Low on Matching Variable

	Right	Wrong	Total
Reference	.2	.1	.3
Focal	.2	.2	.4

High on Matching Variable

	Right	Wrong	Total
Reference	.5	.2	.7
Focal	.4	.2	.6

Now suppose that target tables are needed for the $n_R = 900$, $n_F = 100$ condition. The reference group probabilities are multiplied by 900 and the focal group probabilities are multiplied by 100, producing the following table for use in DIF analysis.

Low on Matching Variable

	Right	Wrong	Total
Reference	180	90	270
Focal	20	20	40

High on Matching Variable

	Right	Wrong	Total
Reference	450	180	630
Focal	40	20	60

Table 1

True and Estimated Item Parameters for Pretest Items

Item	b	d	$\hat{a}$	$\hat{b}$	$\hat{c}$[a]
1	-1.30	-.70	.97	-1.32	.14
2	-1.30	-.35	.90	-1.42	.14
3	-1.30	0	.89	-1.33	.14
4	-1.30	.35	1.06	-1.35	.14
5	-1.30	.70	.93	-1.38	.14
6	0	-.70	1.13	.08	.19
7	0	-.35	1.09	.04	.15
8	0	0	1.04	-.04	.15
9	0	.35	1.23	.06	.19
10	0	.70	1.07	-.07	.13
11	1.30	-.70	1.13	1.26	.16
12	1.30	-.35	1.06	1.38	.16
13	1.30	0	1.21	1.20	.16
14	1.30	.35	1.24	1.35	.17
15	1.30	.70	.92	1.28	.13

[a]In the LOGIST program, estimated c parameters are set to a common value for items on which the c parameter cannot be estimated accurately. This applies to Items 1-5.

Note. All a parameters are 1 and all c parameters are .15.

Table 2

Median Residuals for Three Pretest Matching Variables

	Reference Group (n_R = 1,000)			Focal Group (n_F = 1,000)		
Item	Old	Optimal	Approximate	Old	Optimal	Approximate
1	-.21	-.11	-.12	-.33	-.45	-.21
2	-.19	-.24	-.12	-.25	-.26	-.18
3	-.18	-.23	-.13	-.18	-.15	-.20
4	-.18	-.25	-.11	-.10	-.05	-.21
5	-.20	-.24	-.12	-.03	.09	-.19
6	-.09	-.07	-.13	-.27	-.34	-.19
7	-.17	-.15	-.15	-.30	-.34	-.22
8	-.16	-.11	-.11	-.18	-.19	-.21
9	-.18	-.15	-.14	-.19	-.15	-.19
10	-.20	-.26	-.12	-.12	-.01	-.22
11	-.14	-.04	-.13	-.19	-.22	-.20
12	-.26	-.24	-.14	-.19	-.19	-.20
13	-.14	-.04	-.13	-.18	-.21	-.20
14	-.14	-.15	-.14	-.17	-.15	-.19
15	-.13	-.11	-.13	-.07	-.15	-.22
Median across items	-.18	-.15	-.13	-.18	-.19	-.20

Note. Each residual was computed by subtracting the true score (equation 4) from the matching variable (equations 1-3). The standard deviation of the true score was about 13 for the reference group and 12 for the focal group.

Table 3

Means and Standard Deviations of DIF Statistics

	$n_R = 500, n_F = 500$		$n_R = 900, n_F = 100$		$n_R = 500, n_F = 100$
Statistic	Old	New	Old	New	New
MH D-DIF	-.02 (1.35)	.04 (1.36)	.02 (1.25)	.05 (1.25)	.05 (1.28)
SE(MH D-DIF)	.41 (.04)	.42 (.04)	.64 (.05)	.65 (.05)	.69 (.06)
STD P-DIF x 10	.00 (.88)	.02 (.86)	.00 (.88)	.02 (.86)	.02 (.86)
SE(STD P-DIF) x 10	.33 (.03)	.34 (.02)	.49 (.05)	.49 (.04)	.52 (.04)

Table 4

Correlations of Pretest DIF Statistics and True DIF (*ad*)
for $n_R = 500$, $n_F = 500$.

	Old *MH D-DIF*	New *MH D-DIF*	Old *STD P-DIF*	New *STD P-DIF*	*ad*
Old *MH D-DIF*	1				
New *MH D-DIF*	1.00	1			.
Old *STD P-DIF*	.99	1.00	1		
New *STD P-DIF*	.99	.99	1.00	1	.
ad	.94	.95	.91	.92	1

Note. "Old" statistics are from the ZTW (1993) study; "new" statistics are from the present study.

Table 5

Correlations of Pretest DIF Statistics and True DIF (*ad*)
for $n_R = 900$, $n_F = 100$

	Old *MH D-DIF*	New *MH D-DIF*	Old *STD P-DIF*	New *STD P-DIF*	*ad*
Old *MH D-DIF*	1				
New *MH D-DIF*	1.00	1			
Old *STD P-DIF*	1.00	.99	1		
New *STD P-DIF*	.99	1.00	1.00	1	
ad	.92	.93	.91	.92	1

Note. "Old" statistics are from the ZTW (1993) study; "new" statistics are from the present study.

Table 6

Expected Percent of "C" Results for Old and New Matching Variables

Item	Nominal Status[a]	$n_R = 500, n_F = 500$ Old	$n_R = 500, n_F = 500$ New	$n_R = 900, n_F = 100$ Old	$n_R = 900, n_F = 100$ New	$n_R = 500, n_F = 100$ New
1	C-	97.7	93.0	70.9	58.8	56.6
2	B-	25.7	20.1	13.6	11.9	11.6
3	A	0.0	0.0	0.2	0.1	0.2
4	B+	4.2	8.7	4.6	7.2	7.1
5	C+	79.7	82.1	46.9	50.4	46.3
6	C-	67.1	62.6	21.3	20.4	21.5
7	B-	6.1	5.3	3.3	3.6	4.0
8	A	0.0	0.0	0.1	0.1	0.1
9	B+	2.4	5.4	2.6	3.6	4.0
10	C+	75.8	83.4	37.5	41.3	40.1
11	C-	0.7	1.0	0.8	1.1	1.4
12	B-	0.1	0.2	0.4	0.6	0.8
13	A	0.0	0.0	0.2	0.2	0.2
14	B+	0.3	0.4	0.7	0.7	0.8
15	C+	13.1	11.4	6.0	4.4	5.1

[a]This column gives the ETS DIF category to which the item nominally belongs. A minus sign indicates that the item is easier for the reference group, while a plus signs indicates that it is easier for the focal group.

Table 7

Distribution of Standard Errors of DIF Statistics
with Optimal Matching Variable

	SE(MH D-DIF)		*SE(STD P-DIF)* x 10	
	n_R =900, n_F =100	n_R =500, n_F =100	n_R =900, n_F =100	n_R =500, n_F =100
Lowest	.60	.63	.42	.45
25th %ile	.62	.65	.46	49
Median	.63	.67	.50	.53
75th %ile	.69	.72	.53	.56
Highest	.78	.84	.55	.58

Table 8

Intercorrelation of Standard Error Estimates for DIF Statistics

	1	2	3	4	5	6
1. Mean *SE(MH D-DIF)* over replications	1					
2. ET *SE(MH D-DIF)*	.99	1				
3. Empirical *SD* of *MH D-DIF*	.82	.81	1			
4. Mean *SE(STD P-DIF)* over replications	-.59	-.59	-.51	1		
5. ET *SE(STD P-DIF)*	-.55	-.56	-.42	.98	1	
6. Empirical *SD* of *STD P-DIF*	-.54	-.55	-.14	.77	.81	1

Note. $n_R = 900$, $n_F = 100$. ET estimates are based on 60,000 cases per group.

Table 9

Comparison of Standard Error Estimates for DIF Statistics:
Median and Range over 15 Items

	MH D-DIF		*STD P-DIF* x 10	
	Median	Range	Median	Range
Mean *SE* over replications	.67	.23	.50	.13
ET standard error	.63	.19	.50	.13
Empirical *SD* over replications	.65	.21	.48	.15

Note. $n_R = 900$, $n_F = 100$. ET estimates are based on 60,000 cases per group.

Figure 1
Standard Error of MH D-DIF for Pretest Items
(nR = 500, nF = 500)

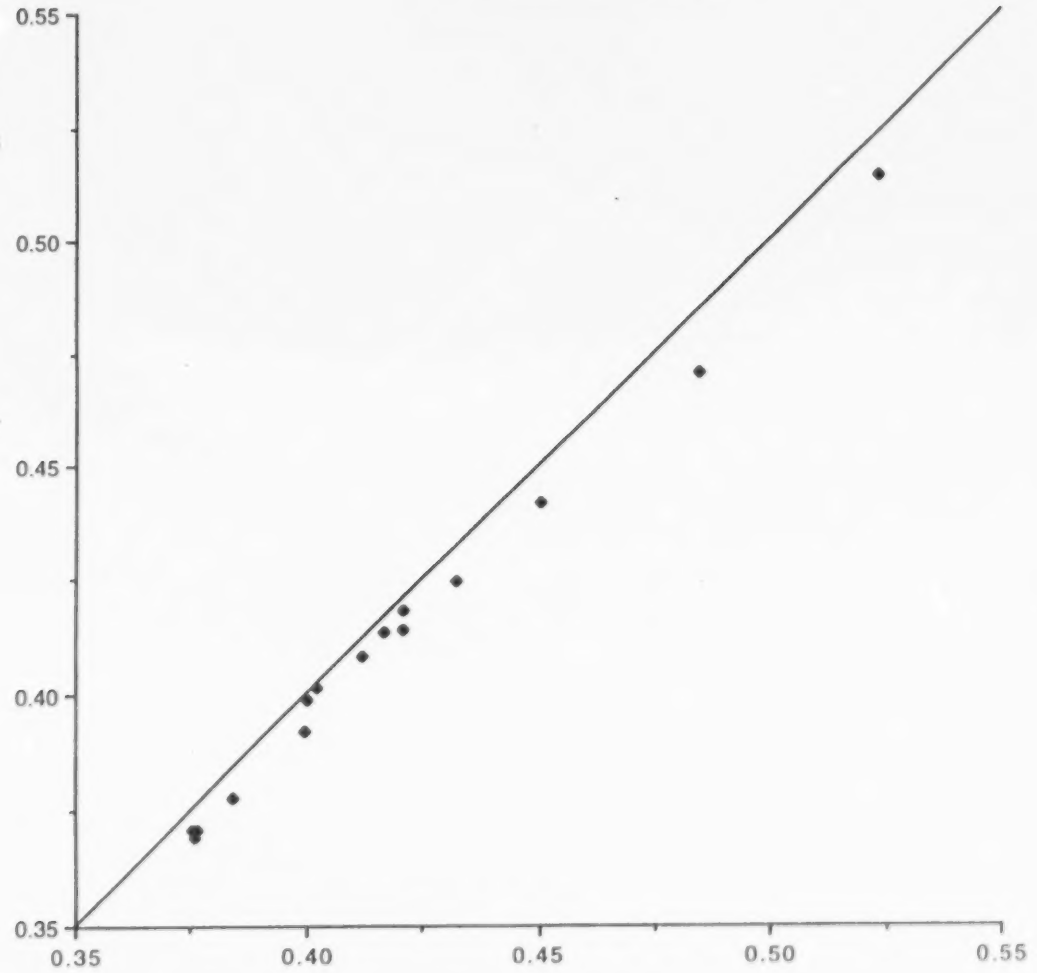


Figure 2
Standard Error of STD P-DIF for Pretest Items
(nR = 500, nF = 500)

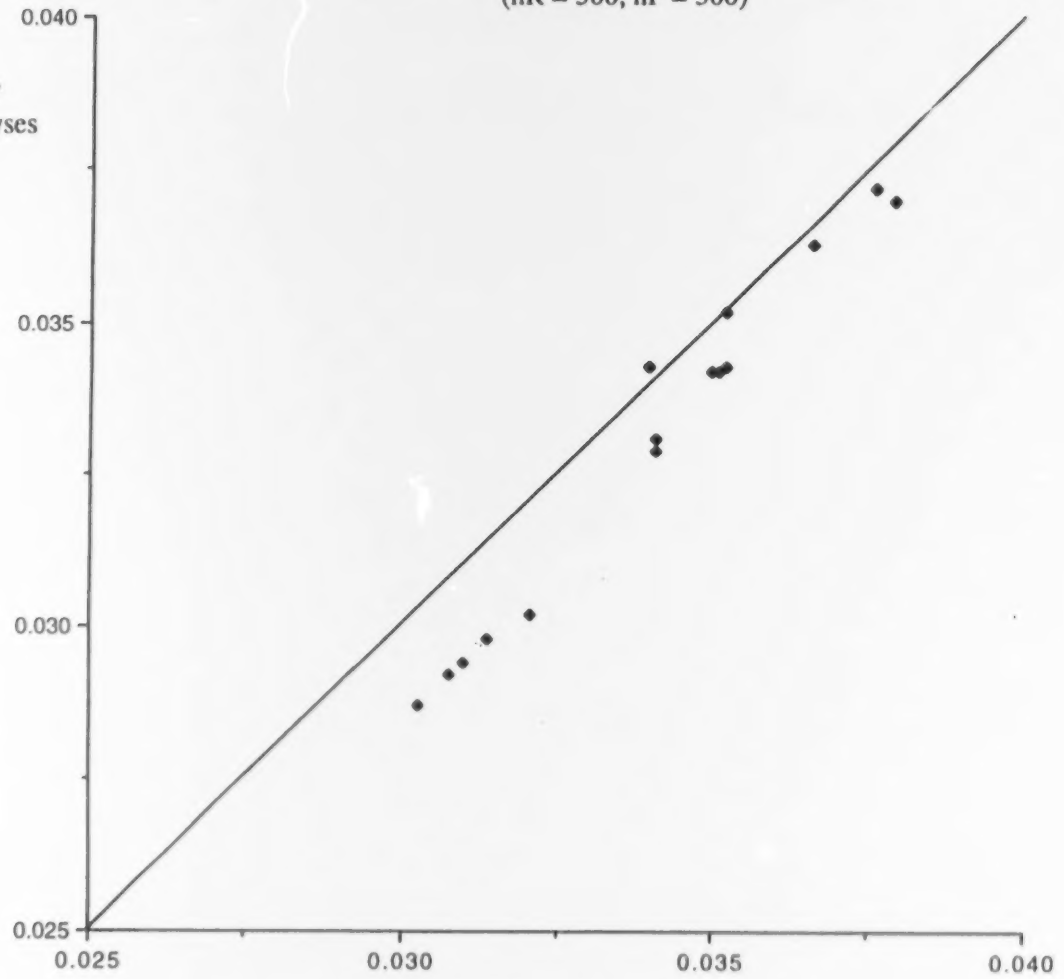


Figure 3
MH D-DIF Values for CAT and Nonadaptive Administration
(nR = 900, nF = 100)

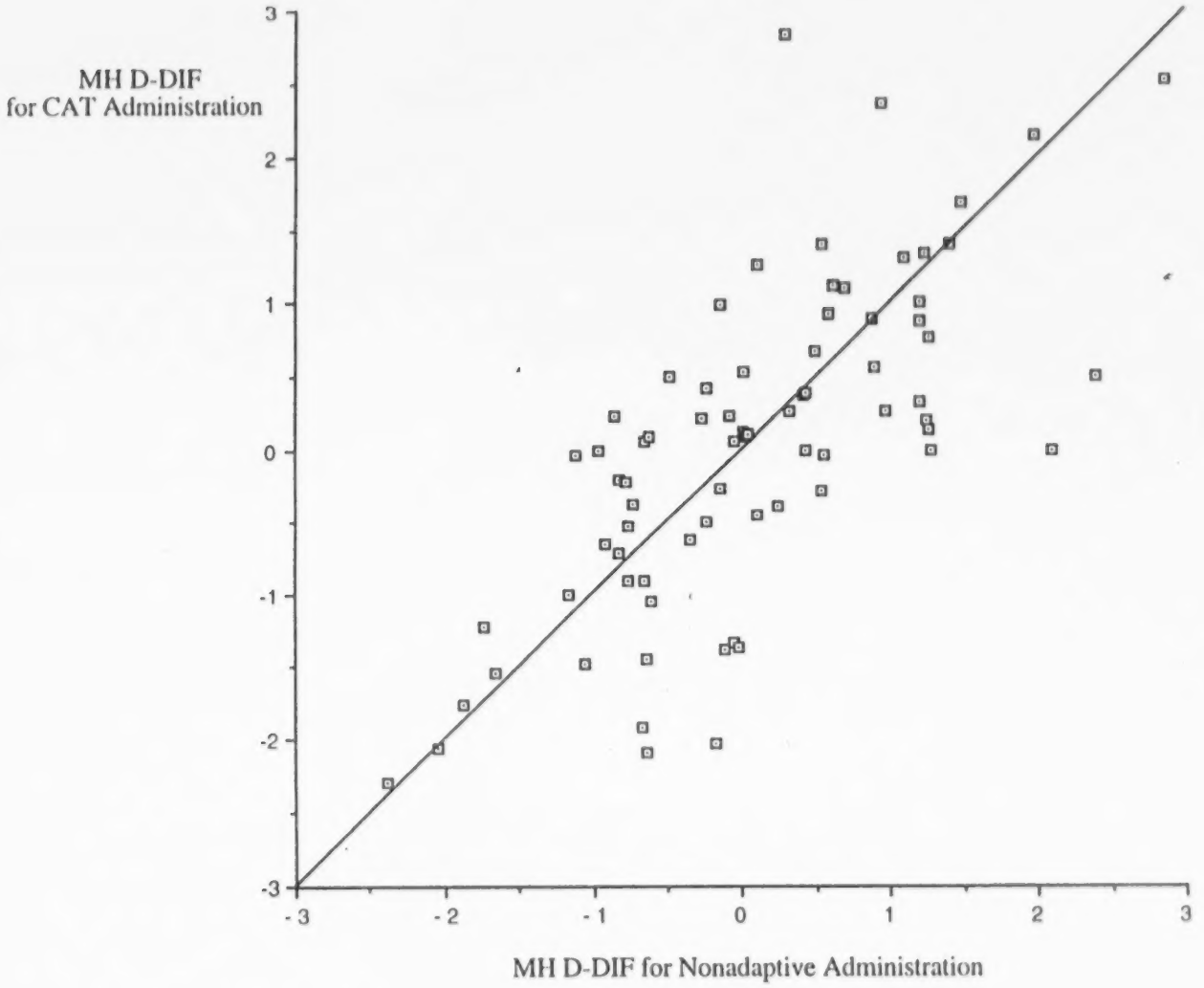
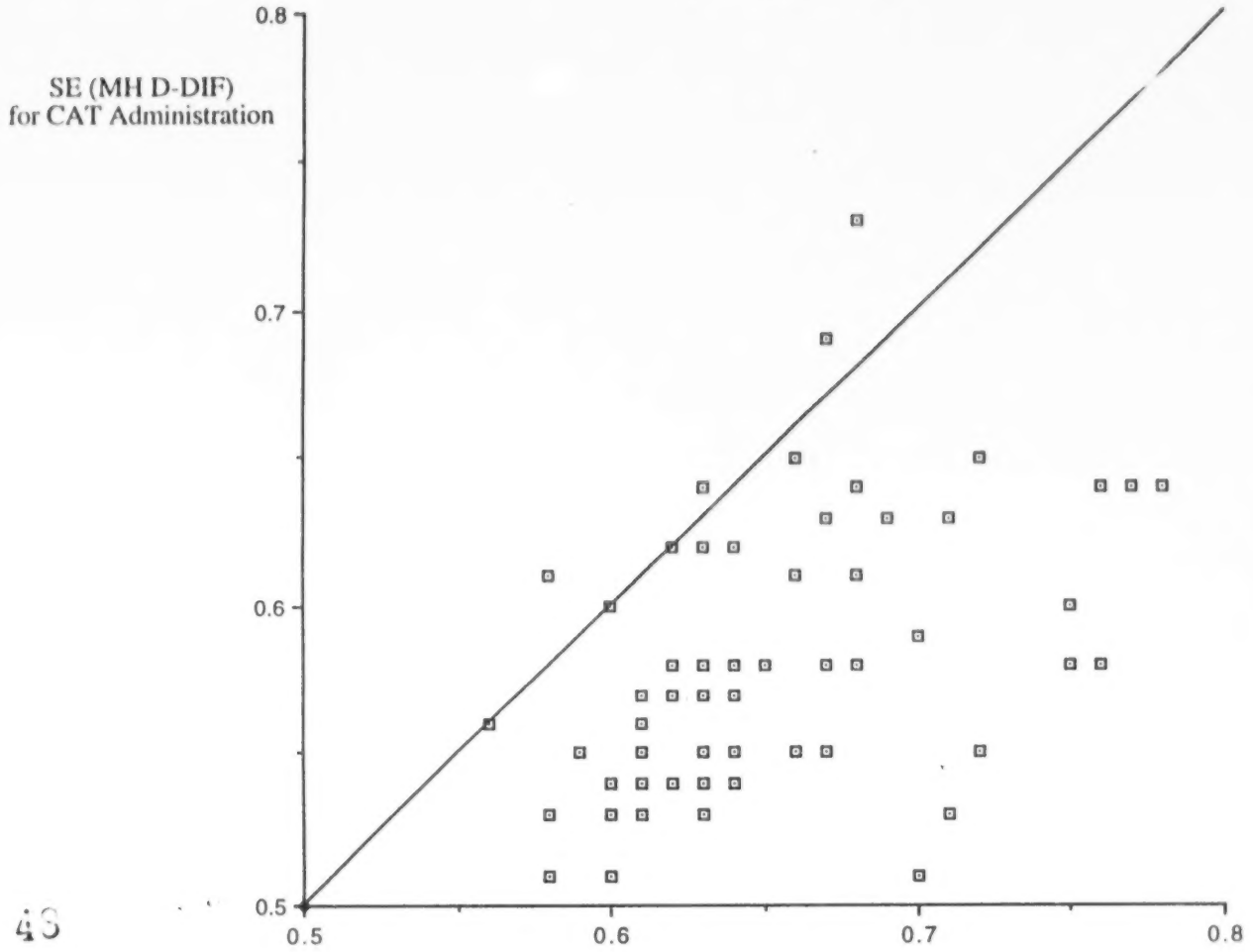


Figure 4
SE (MH D-DIF) Values for CAT and Nonadaptive Administration
(nR = 900, nF = 100)



43

Figure 5
STD P-DIF Values for CAT and Nonadaptive Administration
(nR = 900, nF = 100)

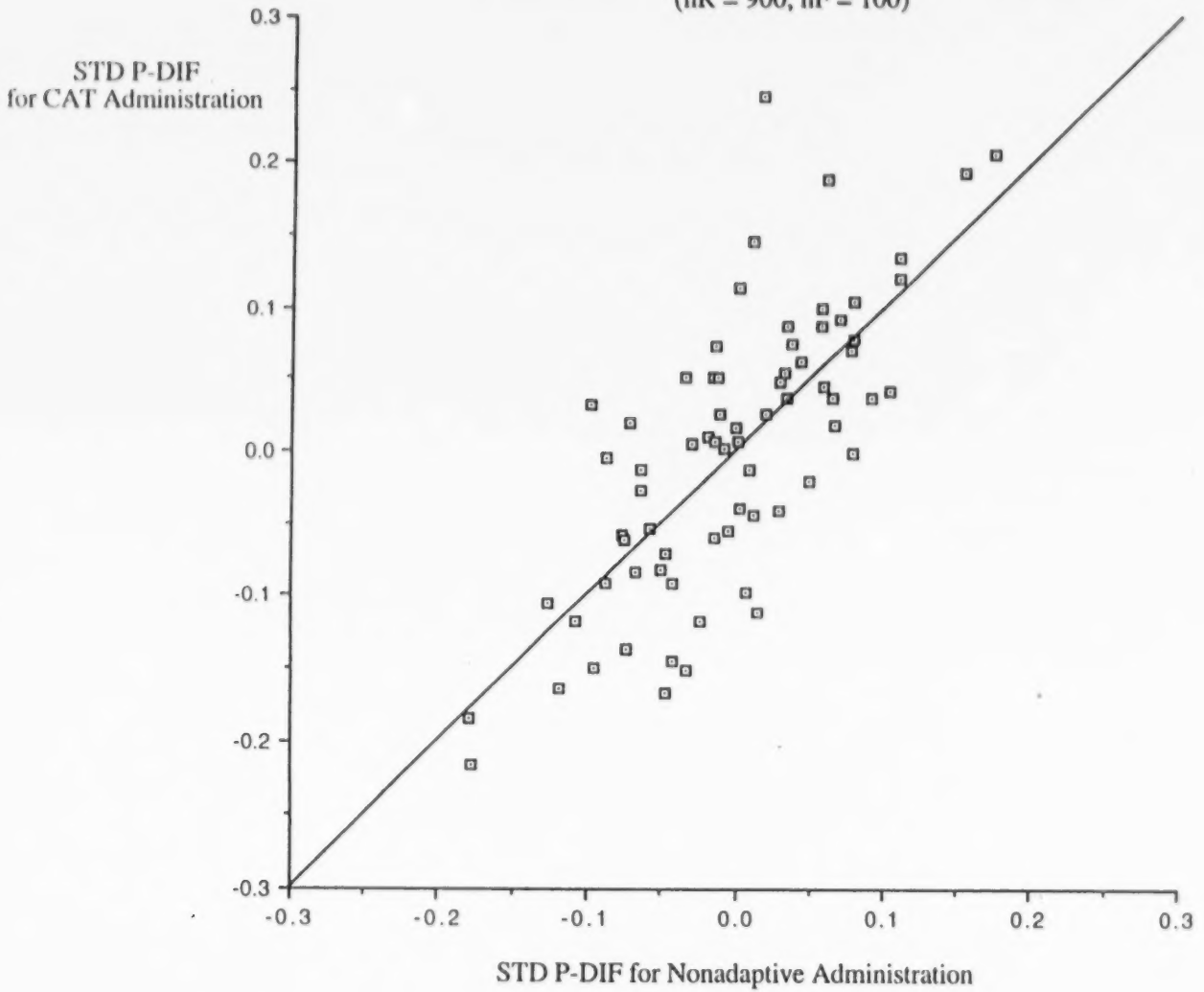


Figure 6
SE (STD P-DIF) Values for CAT and Nonadaptive Administration
(nR = 900, nF = 100)